Check for updates

DATA NOTE

ERGA-BGE reference genome of *Phorcus turbinatus* - a water quality biomarker and species distributed along the Mediterranean coast

[version 1; peer review: 4 approved with reservations]

Judit Salces-Ortiz[1], Klara Eleftheriadi[1], Fernando Ángel Fernandez-Álvarez[2], Rosa Fernández[1], Astrid Böhne[3], Rita Monteiro[3], Thomas Marcussen[4,5], Torsten H. Struck[4], Rebekah A. Oomen[4-8], Marta Gut[9,10], Laura Aguilera[9,10], Francisco Câmara Ferreira[9,10], Fernando Cruz[9,10], Jèssica Gómez-Garrido[9,10], Tyler S. Alioto[9,10], Fergal Martin[11], Anna Lazar[11], Leanne Haggerty[11], Chiara Bortoluzzi[12]

[1]Metazoa Phylogenomics and Genome Evolution Lab, Passeig maritim de la Barceloneta 37-49, Institute of Evolutionary Biology (CSIC-UPF)., Barcelona, 08003, Spain
[2]Centro Oceanográfico de Gijón (IEO, CSIC), Av. del Príncipe de Asturias 70 Bis, Instituto Español de Oceanografía, Gijón, 33212, Spain
[3]Leibniz Institute for the Analysis of Biodiversity Change, Adenauerallee127, Museum Koenig Bonn, Bonn, 53113, Germany
[4]University of Oslo, P.O. Box 1172, Blindern, Natural History Museum, Oslo, 0318, Norway
[5]University of Oslo, Blindernveien 31, Centre for Ecological & Evolutionary Synthesis, Oslo, 0371, Norway
[6]Department of Biological Sciences, 100 Tucker Park Road, University of New Brunswick Saint John, Saint John, E2K5E2, Canada
[7]Tjärnö Marine Laboratory, Hättebäcksvägen 7, University of Gothenburg, Gothenburg, 45296, Sweden
[8]Centre for Coastal Research, Universitetsveien 25, University of Agder, Kristiansand, 4630, Norway
[9]Centro Nacional de Análisis Genómico (CNAG), Baldiri Reixac 4, Barcelona, 08028, Spain
[10]Universitat de Barcelona (UB), Barcelona, 08028, Spain
[11]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, CB10 1SD, UK
[12]SIB Swiss Institute of Bioinformatics, Amphipôle, Quartier UNIL-Sorge, Lausanne, 1015, Switzerland

v1 **First published:** 25 Mar 2026, **6**:79
https://doi.org/10.12688/openreseurope.23162.1
Latest published: 25 Mar 2026, **6**:79
https://doi.org/10.12688/openreseurope.23162.1

Abstract

Phorcus turbinatus (turbinate monodont; Mollusca, Trochidae) is widely used as a biomarker of water quality and numerous scientific studies rely on this species for environmental and genomic research. Consequently, the generation of a chromosome-level assembly for *P. turbinatus* is of substantial scientific interest. Most of the genome sequence was assembled into 18 contiguous chromosomal pseudomolecules and the mitochondrial genome. This chromosome-level assembly encompasses 0.95 Gb, composed of 104 contigs and 103 scaffolds, with a contig and scaffold N50 of 50.6 Mb.

Open Peer Review

Approval Status ? ? ? ?

	1	2	3	4
version 1 25 Mar 2026	? view	? view	? view	? view

1. **André Gomes-dos-Santos**, University of Porto, Porto, Portugal

2. **Xiao-Chen Huang**, Nanchang University, Nanchang, China

Keywords
Phorcus turbinatus, genome assembly, European Reference Genome Atlas, Biodiversity Genomics Europe, Earth Biogenome Project, biodiversity hotspot

3. **Marco Gerdol**, University of Trieste, Trieste, Italy

4. **Masa-Aki Yoshida**, Shimane University, Shiman, Japan

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Horizon Europe gateway.

This article is included in the Genome Reports from the Biodiversity Genomics Europe Project collection.

Corresponding author: Chiara Bortoluzzi (chiara.bortoluzzi@sib.swiss)

Author roles: Salces-Ortiz J: Investigation, Methodology, Project Administration, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Eleftheriadi K**: Investigation, Methodology, Project Administration, Resources, Writing – Review & Editing; **Fernandez-Álvarez FÁ**: Investigation, Methodology, Project Administration, Resources, Writing – Review & Editing; **Fernández R**: Investigation, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing; **Böhne A**: Methodology, Project Administration, Supervision, Writing – Review & Editing; **Monteiro R**: Methodology, Project Administration, Supervision, Writing – Review & Editing; **Marcussen T**: Methodology, Project Administration, Supervision, Writing – Review & Editing; **H. Struck T**: Methodology, Project Administration, Supervision, Writing – Review & Editing; **A. Oomen R**: Methodology, Project Administration, Supervision, Writing – Review & Editing; **Gut M**: Investigation, Writing – Review & Editing; **Aguilera L**: Investigation, Writing – Original Draft Preparation, Writing – Review & Editing; **Câmara Ferreira F**: Data Curation, Formal Analysis, Writing – Review & Editing; **Cruz F**: Data Curation, Formal Analysis, Writing – Review & Editing; **Gómez-Garrido J**: Data Curation, Formal Analysis, Writing – Review & Editing; **S. Alioto T**: Data Curation, Formal Analysis, Supervision, Writing – Review & Editing; **Martin F**: Data Curation, Formal Analysis, Writing – Review & Editing; **Lazar A**: Data Curation, Formal Analysis, Writing – Review & Editing; **Haggerty L**: Data Curation, Formal Analysis, Writing – Review & Editing; **Bortoluzzi C**: Visualization, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: Biodiversity Genomics Europe (Grant no. 101059492) is funded by Horizon Europe under the Biodiversity, Circular Economy and Environment call (REA.B.3); co-funded by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract numbers 22.00173 and 24.00054; and by the UK Research and Innovation (UKRI) under the Department for Business, Energy and Industrial Strategy's Horizon Europe Guarantee Scheme. F.Á.F.-Á. was supported by a Ramón y Cajal fellowship (Ref. RYC2023-043494-I) funded by MCIN/AEI /10.13039/501100011033 and FSE+.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2026 Salces-Ortiz J *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Salces-Ortiz J, Eleftheriadi K, Fernandez-Álvarez FÁ *et al.* **ERGA-BGE reference genome of *Phorcus turbinatus* - a water quality biomarker and species distributed along the Mediterranean coast [version 1; peer review: 4 approved with reservations]** Open Research Europe 2026, **6**:79 https://doi.org/10.12688/openreseurope.23162.1

First published: 25 Mar 2026, **6**:79 https://doi.org/10.12688/openreseurope.23162.1

Introduction

Phorcus turbinatus is a marine gastropod mollusc belonging to the family Trochidae. The species is endemic to Mediterranean coastal regions, particularly along the shores of Spain, Italy and Greece, an area recognised as a European biodiversity hotspot. *P. turbinatus* commonly inhabits intertidal and shallow subtidal rocky habitats, where it contributes to the regulation of algal growth and plays an important role in structuring coastal benthic ecosystems. Owing to its sensitivity to environmental stressors and its ability to bioaccumulate pollutants, *P. turbinatus* is widely used as a biomarker species for assessing water quality and coastal ecosystem health and is frequently included in environmental monitoring and ecological studies.

Despite its scientific relevance, *P. turbinatus* is not currently listed under any conservation category in the IUCN Red List and no specific conservation measures have been reported. Given its ecological importance and extensive use in research, the generation of a chromosome-level genome assembly for *P. turbinatus* represents a valuable resource. Making such genomic data rapidly available to the scientific community will support future studies in environmental monitoring, ecology, and evolutionary biology.

The generation of this reference resource was coordinated by the European Reference Genome Atlas (ERGA) initiative's Biodiversity Genomics Europe (BGE) project, supporting ERGA's aims of promoting transnational cooperation to promote advances in the application of genomics technologies to protect and restore biodiversity (Mazzoni et al., 2023).

Materials & Methods

ERGA's sequencing strategy includes Oxford Nanopore Technology (ONT) and/or Pacific Biosciences (PacBio) for long-read sequencing, along with Hi-C sequencing for chromosomal architecture, Illumina Paired-End (PE) for polishing (i.e. recommended for ONT-only assemblies), and RNA sequencing for transcriptomic profiling, to facilitate genome assembly and annotation. In the case of this species, it was not possible to obtain a Hi-C library. However, chromosome territories could be confirmed by alignment to the chromosome-level assembly of a sister species, *Phorcus lineatus* (xgPhoLine1.1).

Sample and sampling information

On 28th November 2022, Judit Salces Ortiz, Klara Eleftheriadi, Fernando A. Fernández-Álvarez, and Rosa Fernández sampled five specimens of *Phorcus turbinatus* (sex unknown). Sampling took place in Cala de Niu, Sant Feliu de Guixols, Spain. The species was identified by Fernando Ángel Fernández-Álvarez based on morphology through the website Malacological Pictures (https://malacopics.nl/db/index.php).

The biological material collected in Spain and used to generate digital sequences, was retrieved from wildlife taxa regulated by the Spanish Royal Decree 124/2017. No ABS permits were granted as utilization of the material falls under exemption in Article 3(2) of the Spanish Royal Decree 124/2017 and was confirmed by the national authorities. Once collected, specimens were euthanized by being flash frozen in liquid nitrogen and were kept at −70 °C until DNA extraction.

Vouchering information

Physical reference materials for the here sequenced specimen have been deposited in the Museo Nacional de Ciencias Naturales (MNCN) https://www.mncn.csic.es/es under the accession number MNCN-ADN 151.725. Frozen reference tissue material of foot muscle is available from the same individual at the Biobank of the Museo Nacional de Ciencias Naturales (MNCN) https://www.mncn.csic.es/es under the voucher ID MNCN-ADN 151.724.

Data availability

Phorcus turbinatus and the related genomic study were assigned to Tree of Life ID (ToLID) 'xgPhoTurb1' and all sample, sequence, and assembly information are available from the INSDC under the umbrella BioProject PRJEB88157. The sample information is available at the following BioSample accessions: SAMEA114492921, SAMEA114492923, and SAMEA114492925. The genome assembly is accessible from ENA under accession number GCA_965237135.1 and the annotated genome is available through the Ensembl webpage (https://projects.ensembl.org/erga-bge/). Sequencing data produced as part of this project are available from ENA at the following accessions: ERX14228952, ERX14228953, ERX14228954, and ERX14228955. Documentation related to the genome assembly and curation can be found in the ERGA Assembly Report (EAR) document available at https://github.com/ERGA-consortium/EARs/tree/main/Assembly_Reports/Phorcus_turbinatus/xgPhoTurb1. Further details and data about the project are hosted on the ERGA portal at https://portal.erga-biodiversity.eu/data_portal/43341.

Genetic information
The estimated genome size, based on ancestral taxa, is 1.91 Gb. This is a diploid genome with a haploid number of 9 chromosomes (2n = 18). All information for this species was retrieved from Genomes on a Tree (Challis et al., 2023).

DNA/RNA processing
DNA was extracted from the body with a shell using the Blood & Cell Culture DNA Mini Kit (Qiagen) following the manufacturer's instructions. DNA quantification was performed using a Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific), and DNA integrity was assessed using a Genomic DNA 165 Kb Kit (Agilent) on the Femto Pulse system (Agilent). The DNA was stored at +4 °C until sequenced.

RNA was extracted from the body with a shell using an RNeasy Mini Kit (Qiagen) according to the manufacturer's instructions. RNA quantification was performed using the Qubit RNA BR kit, and RNA integrity was assessed using a Bioanalyzer 2100 system (Agilent) RNA 6000 Nano Kit (Agilent). The RNA was stored at −80 °C until sequenced.

Library preparation and sequencing
For long-read whole genome sequencing, two libraries were prepared using the SQK-LSK114 Kit (Oxford Nanopore Technologies, ONT), which were then sequenced across three R10.4.1 flow cells on a PromethION 24 A Series instrument (ONT). A short-read whole-genome sequencing library was prepared using the KAPA Hyper Prep Kit (Roche). The RNA library was prepared using the KAPA mRNA Hyper prep kit (Roche). All short-read libraries were sequenced on a NovaSeq 6000 instrument (2x151bp, Illumina). In total 234x Oxford Nanopore and 135x Illumina WGS shotgun data were sequenced to generate the assembly.

Genome assembly methods
The genome was assembled using the CNAG CLAWS pipeline v2.3 (Gomez-Garrido, 2024). Briefly, reads were preprocessed for quality and length using Trim Galore v0.6.7 and Filtlong v0.2.1, and initial contigs were assembled using hifiasm v0.24.0 (Cheng et al. 2024) with options --dual-scaf --ont. As we used Q20+ ONT and QV of the primary assembly was already 59.7, no polishing with Illumina was required. In absence of Hi-C data, we aligned the contigs to the *Phorcus lineatus* chromosome-level assembly (GCA_921293015.1) using minimap2 v2.24 (Li 2016) with a maximum sequence divergence of 1%. The whole genome alignment (WGA) confirmed chromosomal status for 17 contigs and supported the joining of two other contigs aligning to chromosome 17 into a super-scaffold. Both contigs contain terminal telomeric repeats confirming their orientation and the telomere-to-telomere integrity of the super-scaffold. Finally, the mitochondrial genome was assembled as a single circular contig of 17,345 bp using the FOAM pipeline (https://github.com/cnag-aat/FOAM) and included in the released assembly (GCA_965237135.1). Summary analysis of the released assembly was performed using the ERGA-BGE Genome Report ASM Galaxy workflow (https://doi.org/10.48546/workflowhub.workflow.1103.2).

Genome annotation methods
A gene set was generated using the Ensembl Gene Annotation system (Aken et al., 2016), primarily by aligning publicly available short-read RNA-seq data from BioSample: SAMEA114492923 to the genome. Gaps in the annotation were filled via protein-to-genome alignments of a select set of clade-specific proteins from UniProt (The UniProt Consortium, 2019), which had experimental evidence at the protein or transcript level. At each locus, data were aggregated and consolidated, prioritising models derived from RNA-seq data, resulting in a final set of gene models and associated non-redundant transcript sets. To distinguish true isoforms from fragments, the likelihood of each open reading frame (ORF) was evaluated against known [metazoan|plant|fungal] proteins. Low-quality transcript models, such as those showing evidence of fragmented ORFs, were removed. In cases where RNA-seq data were fragmented or absent, homology data were prioritised, favouring longer transcripts with strong intron support from short-read data. The resulting gene models were classified into two categories: protein-coding, and long non-coding. Models that did not overlap protein-coding genes and were constructed from transcriptomic data were considered potential lncRNAs. Potential lncRNAs were further filtered to remove single-exon loci due to their unreliability. Putative miRNAs were predicted by performing a BLAST search of miRBase (Kozomara et al., 2019) against the genome, followed by RNAfold analysis (Gruber et al., 2008). Other small non-coding loci were identified by scanning the genome with Rfam (Kalvari et al., 2018) and passing the results through Infernal (Nawrocki & Eddy, 2013). Summary analysis of the released annotation was carried out using the ERGA-BGE Genome Report ANNOT Galaxy workflow (https://doi.org/10.48546/workflowhub.workflow.1096.1).

Results
Genome assembly
The genome assembly has a total length of 951,702,898 bp in 103 scaffolds including the mitogenome (Figures 1 & 2), with a GC content of 35.5%. The assembly has a contig N50 of 50,589,882 bp and L50 of 8 and a scaffold N50 of

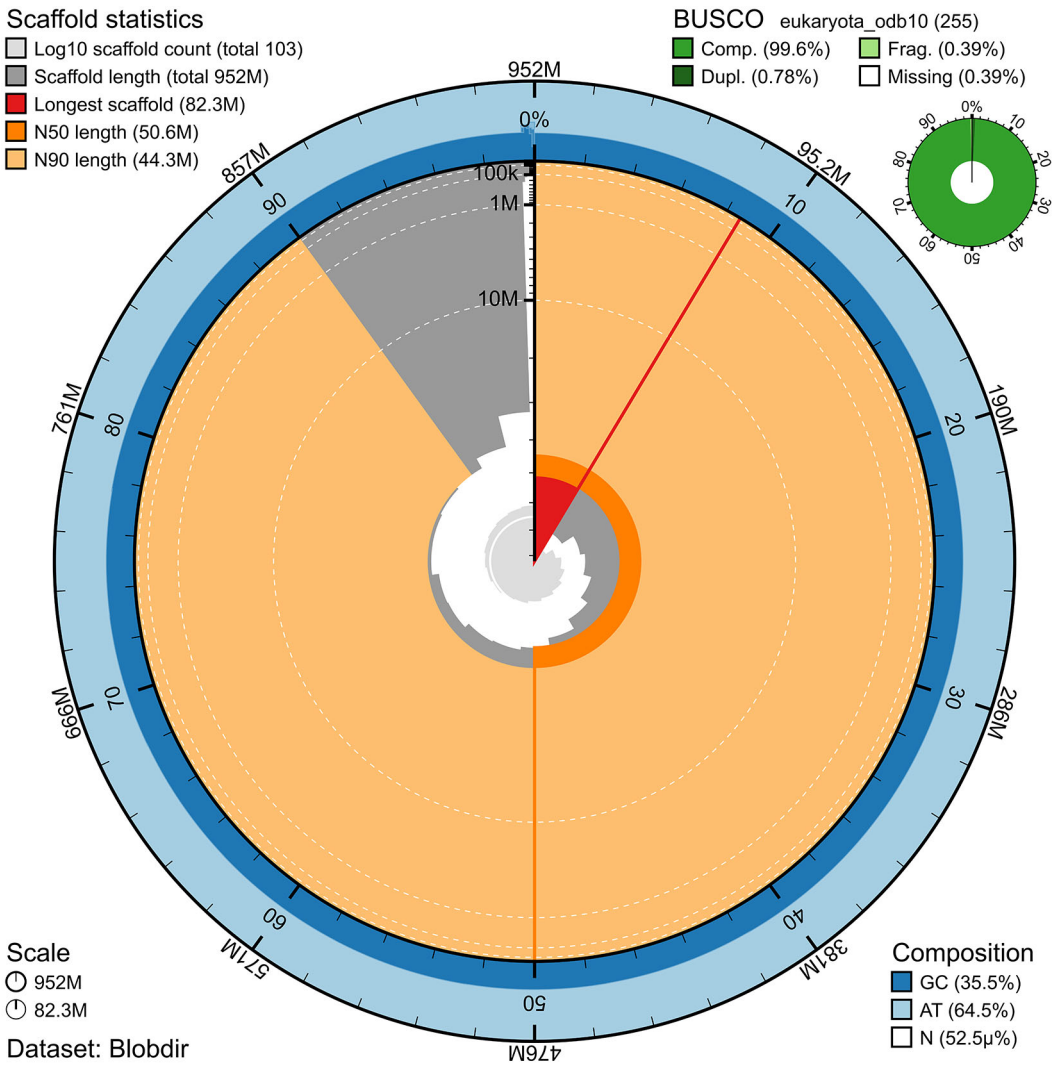


Figure 1. Snail plot summary of assembly statistics. The main plot is divided into 1,000 size-ordered bins around the circumference, with each bin representing 0.1% of the 951,702,898 bp assembly including the mitochondrial genome. The distribution of sequence lengths is shown in dark grey, with the plot radius scaled to the longest sequence present in the assembly (82.3 Mb, shown in red). Orange and pale-orange arcs show the scaffold N50 and N90 sequence lengths (50,589,882 and 44,257,409 bp), respectively. The pale grey spiral shows the cumulative sequence count on a log-scale, with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT, and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated, and missing BUSCO genes found in the assembled genome from the Eukaryota database (odb10) is shown in the top right.

50,589,882 bp and L50 of 8. The assembly has a total of one gap, totalling 0.5 kb in cumulative size. The total sequence assigned to chromosomes is 99.57%. The divergence time between *P. turbinatus* and sister species *P. lineatus* was estimated to be as high as 16 MYA (14.0–16.1 MYA) (Kumar et al. 2017). Despite the large divergence, chromosomes are strongly collinear, showing several inversions but few translocations (Figure 2). The single-copy gene content analysis using the Eukaryota database (odb10) with BUSCO (Manni et al., 2021) resulted in 99.6% completeness (98.9% single and 0.8% duplicated). 64.74% of reads k-mers were present in the assembly and the assembly has a base accuracy Quality Value (QV) of 73.83 as calculated by Merqury (Rhie et al., 2020).

Genome annotation

The genome annotation consists of 40,094 protein-coding genes with associated 58,414 transcripts, in addition to 10,944 non-coding genes (Table 1). Using the longest isoform per transcript, the single-copy gene content analysis using the Eukaryota database (odb10) with BUSCO resulted in 99.2% completeness. Using the OMAmer Metazoa-v2.0.0.h5 database for OMArk (Nevers et al., 2025) resulted in 96.0% completeness and 62.0% consistency (Table 2).

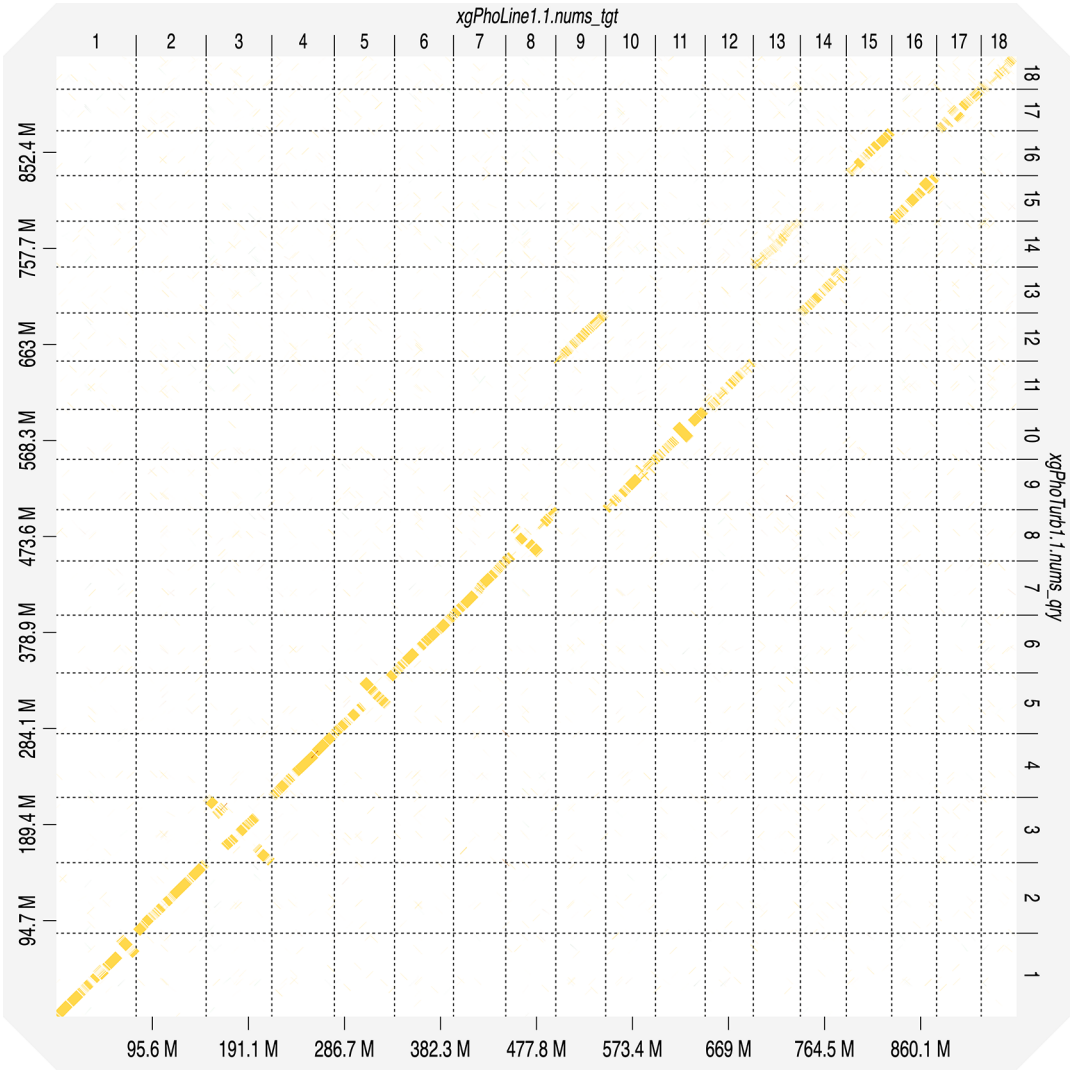


Figure 2. Whole-genome-alignment with *Phorcus lineatus*. Dot-Plot of the minimap2 WGA between the two chromosome-level assemblies of *Phorcus lineatus* (X-axis) and *Phorcus turbinatus* (Y-axis) showing the strongest alignments between chromosomes. The plot was produced using D-GENIES (Cabanettes and Klopp, 2018). Note that chromosomes are numbered according to size, not karyotype; for example, chr12 of *P. turbinatus* and chr9 of *P.lineatus* derive from the same ancestral chromosome but have different names.

Table 1. Statistics from assembled gene models.

	No. genes	No. transcripts	Mean gene length (bp)	No. single-exon genes	Mean exons per transcript
mRNA	40,094	58,414	16,269	1,248	9.3
pseudogene	0.0	0.0	0.0	0.0	0.0
snoRNA	98	98	150	98	1.0
lncRNA	3,741	4,101	7,886	15	2.4
tRNA	6,007	6,007	78	6,007	1.0

Table 1. *Continued*

	No. genes	No. transcripts	Mean gene length (bp)	No. single-exon genes	Mean exons per transcript
snRNA	969	969	159	969	1.0
rRNA	128	128	532	128	1.0
scRNA	1	1	129	1	1.0
Other ncRNA	10,979	10,979	75–105	10,979	1.0–1.0

Table 2. Annotation completeness and consistency scores calculated by BUSCO run in protein mode (eukaryota_odb10) and OMArk (Metazoa-v2.0.0.h5).

	Complete	Singular	Duplicated	Fragmented	Missing
BUSCO	253 (99.2%)	0.0 (0.0%)	253 (99.2%)	1 (0.4%)	1 (0.4%)
OMArk	2,065 (96.0%)	346 (16.1%)	1,719 (79.9%)	-	88 (4.1%)
	Consistent	**Inconsistent**	**Contaminants**	**Unknown**	
OMArk	24,868 (62.0%)	4,478 (11.2%)	0.0 (0.0%)	10,748 (26.8%)	

Author contributions

JS-O, KE, FÁFÁ and RF coordinated the project; JS-O, KE, FÁFÁ and RF collected the species; FÁFÁ identified the species; JS-O sampled and preserved biological material and provided metadata; AsB, RM, TM, RAO, and THS provided support in sampling, shipping of biological material, metadata collection, and management; LA and MG extracted DNA, prepared libraries, and performed sequencing; FCF, FC, and JG-G performed genome assembly and curation under the supervision of TSA; FM and LH performed genome annotation; CB generated the analysis and report. All authors contributed to the writing, review, and editing of this genome note and read and approved the final version.

Acknowledgements

We acknowledge the support of the Freiburg Galaxy Team: Saim Momin and Björn Grüning, Bioinformatics, University of Freiburg (Germany), funded by the German Federal Ministry of Education and Research BMBF grant 031 A538A de. NBI-RBC and the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) within the framework of LIBIS/de. NBI Freiburg. We would like to acknowledge the assembly reviewer Benjamin Istace from Genoscope.

References

Aken BL, Ayling S, Barrell D, *et al.*: **The Ensembl gene annotation system.** *Database.* 2016; **2016**: baw093.
Publisher Full Text

Cabanettes F, Klopp C: **D-GENIES: dot plot large genomes in an interactive, efficient and simple way.** *PeerJ.* 2018; **6**: e4958.
PubMed Abstract | Publisher Full Text | Free Full Text

Challis R, Kumar S, Sotero-Caio C, *et al.*: **Genomes on a Tree (GoaT): A versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic tree of life.** *Wellcome Open Research.* 2023; **8**: 24.
Publisher Full Text

Cheng H, Asri M, Lucas J, *et al.*: **Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph.** *Nat. Methods.* 2024; **21**: 967–970.
PubMed Abstract | Free Full Text

Consortium, U: **UniProt: A worldwide hub of protein knowledge.** *Nucleic Acids Res.* 2019; **47**(D1): D506–D515.

Gomez-Garrido J: **CLAWS (CNAG's Long-read Assembly Workflow in Snakemake) [Computer software].** 2024.

Gruber AR, Lorenz R, Bernhart SH, *et al.*: **The vienna RNA websuite.** *Nucleic Acids Res.* 2008; **36**(suppl_2): W70–W74.
PubMed Abstract | Publisher Full Text | Free Full Text

Kalvari I, Nawrocki EP, Argasinska J, *et al.*: **Non-Coding RNA Analysis Using the Rfam Database.** *Curr. Protoc. Bioinformatics.* 2018; **62**(1): e51.
PubMed Abstract | Publisher Full Text | Free Full Text

Kozomara A, Birgaoanu M, Griffiths-Jones S: **miRBase: From microRNA sequences to function.** *Nucleic Acids Res.* 2019; **47**(D1): D155–D162.
Publisher Full Text

Kumar S, Stecher G, Suleski M, *et al.*: **TimeTree: a resource for timelines, timetrees, and divergence times.** *Mol. Biol. Evol.* 2017; **34**(7): 1812–1819.
PubMed Abstract | Publisher Full Text

Li H: **Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences.** *Bioinformatics.* 2016; **32**(14): 2103–2110.
Publisher Full Text

Manni M, Berkeley MR, Seppey M, *et al.*: **BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes.** *Mol. Biol. Evol.* 2021; **38**(10): 4647–4654.
PubMed Abstract | Publisher Full Text | Free Full Text

Mazzoni CJ, Ciofi C, Waterhouse RM: **Biodiversity: An atlas of European reference genomes.** *Nature.* 2023; **619**: 252–252.
Publisher Full Text

Nawrocki EP, Eddy SR: **Infernal 1.1: 100-fold faster RNA homology searches.** *Bioinformatics.* 2013; **29**(22): 2933–2935.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Nevers Y, Warwick Vesztrocy A, Rossier V, *et al.*: **Quality assessment of gene repertoire annotations with OMArk.** *Nat. Biotechnol.* 2025; **43**(1): 124–133.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Rhie A, Walenz BP, Koren S, *et al.*: **Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Peer Review Status: ? ? ? ?

Version 1

Reviewer Report 06 July 2026

https://doi.org/10.21956/openreseurope.25061.r75562

© **2026 Yoshida M.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Masa-Aki Yoshida
Shimane University, Shiman, Japan

The authors present a highly contiguous genome assembly of *Phorcus turbinatus*, a species of ecological and environmental relevance. The underlying sequence assembly looks high quality, supported by substantial long- and short-read coverage, a contig N50 of approximately 50.6 Mb, high BUSCO completeness, and a high Merqury QV. I therefore do not have major concerns regarding the basic report of sequencing data or the overall sequence-level quality of the assembly.

My main concerns relate instead to the interpretation and validation of the chromosome-scale assembly, the internal consistency of the reported genome characteristics, and the apparent redundancy of the genome annotation. These issues are important for a reference genome report, but I believe that they should be resolvable largely by clarifying the analytical procedures, reconciling the reported statistics, and, where necessary, reanalysing the existing data. I would be supportive of indexing once the following points are satisfactorily addressed.

Major Points:

1. Non-independent validation of chromosome-scale pseudomolecules

In the absence of Hi-C data, the authors used the chromosome-level assembly of the sister species *Phorcus lineatus* to assign *P. turbinatus* contigs to chromosomes and, in one case, to support the joining of two contigs into a super-scaffold. The same *P. lineatus* assembly is then used in Figure 2 to demonstrate chromosome-scale collinearity between the two species. This raises an important issue regarding the independence of the validation. Because the cross-species alignment contributed directly to the construction of the *P. turbinatus* pseudomolecules, the resulting dot plot cannot, by itself, be regarded as independent confirmation of their chromosome-scale organization. In particular, the agreement between the final pseudomolecules and the *P. lineatus* chromosomes is at least partly expected from the assembly procedure itself.

I do not consider additional Hi-C sequencing to be necessarily required for this Data Note.

However, the authors should clearly distinguish between evidence used to construct the pseudomolecules and evidence used to validate them. Figure 2 should not be presented as independent validation of chromosome-scale structure unless additional independent support is provided.

The authors should also present more explicit evidence supporting the proposed pseudomolecule structures, particularly for the super-scaffold generated by joining two contigs. For example, the manuscript should show or clearly describe the positions of terminal telomeric repeats, the evidence supporting contig orientation, and any long-read, assembly-graph, or other independent support for the scaffold join. It would be appropriate to include a dot plot drawn using the original scaffolds, without the super-scaffolding using P. lineatus, as a supplementary figure. The interpretation of Figure 2 should be revised accordingly.

2. Internal inconsistency in chromosome number and ploidy

The manuscript contains an apparent inconsistency between the reported chromosome number and the structure of the released assembly. The Genetic Information section states that *P. turbinatus* has a haploid chromosome number of 9 and a diploid chromosome number of 18 ("haploid number of 9 chromosomes [2n = 18]"). However, the Abstract reports 18 chromosomal pseudomolecules, and Figure 2 also shows 18 *P. turbinatus* chromosomes.

For a haploid reference assembly, a species with n = 9 would normally be expected to yield nine chromosome-scale pseudomolecules. Conversely, if the assembly truly represents 18 distinct chromosomes, the reported haploid and diploid chromosome numbers require correction or further explanation. This issue is particularly important because the assembly itself shows very low BUSCO duplication, which does not obviously support the interpretation that the 18 pseudomolecules represent two separately assembled haplotypes.

The authors should therefore carefully verify and reconcile the following points: the cytogenetically established chromosome number of *P. turbinatus*, whether the released assembly is intended to represent a haploid or diploid genome, and why the assembly contains 18 chromosomal pseudomolecules. Any incorrect statement should be corrected consistently throughout the Abstract, Methods, Results, figures, and associated metadata.

3. Discrepancy between estimated genome size, assembly size, and k-mer completeness

The relationship between the estimated genome size and the final assembly size requires substantial clarification. The manuscript reports an estimated genome size of 1.91 Gb, whereas the released assembly is approximately 0.95 Gb, almost exactly half of the reported estimate. At the same time, Merqury indicates that only 64.74% of read k-mers are represented in the assembly, despite the very high BUSCO completeness of 99.6%.

These observations may have a valid explanation, but they are currently difficult to reconcile. High BUSCO completeness indicates that most conserved gene space has been recovered, but it does not by itself demonstrate completeness of the entire genome, particularly repetitive or highly heterozygous regions. The relatively low k-mer completeness therefore deserves explicit discussion.

The authors should clarify whether the reported 1.91-Gb value represents a haploid or diploid genome-size estimate, precisely how this estimate was obtained, and how the approximately two-fold discrepancy between the estimated genome size and the final assembly size should be interpreted. They should also explain the relatively low Merqury k-mer completeness value and discuss whether it reflects heterozygosity, repetitive sequence collapse, differences in genome-size definitions, or another factor.

Because substantial Illumina whole-genome sequencing data are already available, a direct k-mer-based estimate of genome size and heterozygosity from these reads would be highly informative and would help resolve this issue without requiring additional sequencing.

4. Extensive duplication and possible redundancy in the genome annotation

The annotation quality metrics show an unusually strong discrepancy between the genome assembly and the annotated gene set. At the assembly level, BUSCO reports only 0.8% duplicated genes. In contrast, the annotation BUSCO analysis reports that all 253 complete BUSCO genes are duplicated, with no complete single-copy BUSCO genes. OMArk similarly reports a very high proportion of duplicated genes. This pattern raises concern about possible redundancy in the annotation, the treatment of alternative transcript isoforms, the input used for BUSCO and OMArk analyses, or a reporting error in Table 2. The issue is particularly important because the manuscript states that the analysis was performed using “the longest isoform per transcript,” which is conceptually unclear; normally, one representative or longest transcript would be selected per gene.

The authors should therefore verify the annotation quality assessment and clearly state the exact input used for BUSCO and OMArk. In particular, they should confirm whether only one representative protein sequence per gene was analysed. If multiple transcript isoforms were included, the analyses should be repeated using one representative protein per gene. If the reported duplication values are correct, the authors should explain the biological or technical basis for this extensive duplication and assess whether the released gene set contains substantial redundancy. The wording “the longest isoform per transcript” should also be corrected and clarified.

Minor Points:

1. Clarification of specimens, BioSamples, sequencing datasets, and voucher materials:

The relationships among the sampled specimens, BioSamples, sequencing datasets, and voucher materials are currently unclear. Five specimens were collected, whereas three BioSample accessions and four sequencing accessions are listed, and the manuscript subsequently refers to “the sequenced specimen” in the singular. Please provide a clear mapping among the individual specimen(s), BioSample accessions, sequencing data types (ONT, Illumina WGS, and RNA-seq), and voucher materials. It should also be explicitly stated whether the ONT and Illumina WGS data used for genome assembly and Merqury evaluation were obtained from the same individual. The description “body with a shell” should also be clarified by specifying the actual tissue or biological material used for DNA and RNA extraction.

2. Use of a lineage-specific BUSCO dataset

The BUSCO analyses were performed only with the broad Eukaryota dataset (eukaryota_odb10; 255 markers). Because *P. turbinatus* is a mollusc, the authors should additionally assess both the genome assembly and the annotated gene set using a more lineage-specific BUSCO dataset, preferably a Mollusca dataset. The current Eukaryota-level analysis may be retained for comparison, but a lineage-specific analysis would provide a substantially more informative assessment of completeness and duplication. This analysis may also help clarify whether the extreme duplication observed in the current annotation BUSCO results persists with a more taxonomically appropriate dataset.

3. Discrepancy between the reported primary-assembly and final-assembly QV values

The manuscript reports a QV of 59.7 for the primary assembly and states that no Illumina polishing was therefore performed, whereas the final assembly is reported to have a Merqury QV of 73.83. Please explain how the QV increased between the primary and final assemblies despite the absence of polishing, and specify whether the two estimates were calculated using the same read dataset, k-mer parameters, and methodology. If differences in curation, haplotig removal, contamination filtering, or other processing steps account for this increase, these steps should be explicitly described.

4. Internal inconsistencies in the annotation summary and Table 1

The summary of non-coding gene annotations is internally inconsistent. The Results state that the annotation contains 10,944 non-coding genes, which corresponds exactly to the sum of the snoRNA, lncRNA, tRNA, snRNA, rRNA, and scRNA categories in Table 1. However, Table 1 additionally lists 10,979 “Other ncRNA” genes, which would bring the total number of non-coding genes to 21,923. Please reconcile these values and clarify which categories are included in the number reported in the text.

In addition, the “Other ncRNA” row reports ranges under columns labelled as means. This should be corrected, or the underlying RNA categories should be presented separately.

5. Unresolved placeholder in the annotation methods

The annotation methods contain an unresolved placeholder: “known [metazoan|plant|fungal] proteins.” Please replace this with the actual protein set used for annotation and specify the relevant taxonomic scope, database version, and selection criteria.

6. Methodological details required for reproducibility

Please provide sufficient methodological detail to make the assembly and quality-assessment procedures reproducible. In particular, the manuscript should report the versions of BUSCO, Merqury, and OMArk and the relevant database versions; the k-mer size and read dataset used for the Merqury analyses; the minimap2 preset and relevant command-line parameters used for the whole-genome alignment; and the method, motif, and criteria used to identify terminal telomeric repeats.

Because telomeric repeats are also used as evidence for the orientation and integrity of the

scaffold generated from two contigs, the telomere-detection procedure is particularly important and should be described explicitly.

7. References supporting the ecological and biomonitoring relevance of the species

Several statements in the Introduction regarding the ecological role of *P. turbinatus*, its contribution to the regulation of algal growth and benthic ecosystem structure, its sensitivity to environmental stressors, its capacity to bioaccumulate pollutants, and its widespread use as a water-quality biomarker are currently unsupported by references. Please provide appropriate citations, particularly because these claims form the main rationale for the ecological and environmental relevance of this reference genome.

Typographical and editorial errors:

The manuscript contains a number of typographical, grammatical, and formatting issues. The authors should carefully proofread the manuscript throughout. Examples include the following:

Affiliations and author information
- "(CSIC-UPF).," should be corrected to "(CSIC-UPF),".
- "Adenauerallee127" should be corrected to "Adenauerallee 127".
- The spelling of "Fernandez-Álvarez" and "Fernández-Álvarez" should be standardised throughout.

Introduction and sample description
- "*P. turbinatusis*" should be corrected to "*P. turbinatus* is" wherever it occurs.
- The phrase "promoting transnational cooperation to promote advances" is repetitive and should be rephrased.
- In "The biological material collected in Spain and used to generate digital sequences, was retrieved…", the comma after "sequences" should be removed.
- "euthanized by being flash frozen" would be more appropriately written as "euthanized by flash freezing".
- "the here sequenced specimen" should be revised, for example to "the sequenced specimen".

Sequencing and assembly methods
- "2x151bp" should be formatted as "2 × 151 bp".
- "In total 234x Oxford Nanopore…" should be revised to "In total, 234× Oxford Nanopore…".
- "As we used Q20+ ONT and QV of the primary assembly…" is grammatically unclear and should be revised.
- "In absence of Hi-C data" should be corrected to "In the absence of Hi-C data".
- Citation formatting should be standardised, for example "Cheng et al. 2024" and "Li 2016".
- The unresolved placeholder "[metazoan|plant|fungal]" should be replaced with the actual taxonomic group used.
- "classified into two categories: protein-coding, and long non-coding" contains an unnecessary comma and should also be checked for consistency with the broader range of non-coding RNA categories reported in Table 1.

Results
- Citation formatting should be standardised, for example "Kumar et al. 2017".

- "64.74% of reads k-mers" should be corrected to "64.74% of read k-mers" or "64.74% of k-mers derived from the reads".
- "base accuracy Quality Value" should be revised for clarity, for example to "base-level accuracy quality value".
- "Using the longest isoform per transcript" is conceptually unclear and should be corrected, most likely to "using the longest isoform per gene", if that is what was actually done.

Figure 1

- The label "Dataset: Blobdir" appears to be a pipeline-generated placeholder or output artefact and should be replaced with an informative dataset or assembly identifier.

Figure 2

- "Whole-genome-alignment" should be corrected to "Whole-genome alignment".
- "Dot-Plot" should be standardised to "dot plot".
- "P.lineatus" should be corrected to "*P. lineatus*".
- The axis labels appear to show internal file names (xgPhoLine1.1.nums_tgt and xgPhoTurb1.1.nums_qry) and should be replaced with clear species or assembly names.
- The statement that two chromosomes "derive from the same ancestral chromosome" may be stronger than warranted unless an explicit ancestral chromosome reconstruction was performed. A more neutral formulation, such as "represent corresponding homologous chromosomes", may be preferable.

Tables

- In Table 1, pseudogene counts should be reported as integers rather than "0.0".
- In Table 1, the "Other ncRNA" row should not report ranges in columns labelled as means unless the underlying categories are explicitly separated or explained.
- In Table 2, "Singular" should be replaced with the standard BUSCO term "Single-copy", unless the terminology is inherited directly from another analysis tool and is explicitly defined.

References

- "Consortium, U:" should be corrected to "The UniProt Consortium:".
- "The vienna RNA websuite." should be corrected to "The Vienna RNA Websuite."

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Partly

Are sufficient details of methods and materials provided to allow replication by others?

Partly

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genome Biology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 26 June 2026

https://doi.org/10.21956/openreseurope.25061.r75564

© **2026 Gerdol M.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Marco Gerdol
University of Trieste, Trieste, Italy

The manuscript by Salces-Ortiz and colleagues reports the genome assembly of Phorcus turbinatus, a marine gastropod mollusk endemic to the Mediterranean Sea. The genome assembly was generated using gold standard methodologies and this project is part of the ERGA network.

Introduction
The claim "*P. turbinatus* is widely used as a biomarker species for assessing water quality and coastal ecosystem health and is frequently included in environmental monitoring and ecological studies" should be substantiated by a couple of appropriate references.

Methods:
The first paragraph is very generic and simply reports the general strategy recommended for sequencing the genomes part of the ERGA initiative. This sentence is not necessary, so please only focus on the library preparation and sequencing strategy that was actually used for this specific project.

"DNA was extracted from the body with a shell" is unclear. The shell was most certainly removed prior to DNA extraction, so I guess the authors are referring to something along the lines of "the whole body (i.e. soft tissues) of a single adult individual". Did not the authors remove the digestive gland? This is usually done for mollusks, as gut contents can lead to unwanted contamination by exogenous DNA.

Genome assembly methods: it is actually unclear whether the Illumina library was used at all during the assembly process, or whether its generation was simply planned for polishing. If this was the case and no polishing was performed (i.e. if the assembly was obtained by only using ONT reads), then the mentions to the Illumina WGS shotgun library could be removed.

"And supported the joining of two other contigs aligning to chromosome 17 into a super-scaffold." Although this is reasonable, events of chromosome fissions between congeneric species are not uncommon. This conclusion could be better supported if the karyotype of the species had been previously observed to be identical in these two species.

Define “clade-specific proteins from UniProt”. Are these Mollusca-specific, Gastropoda-specific, or else?

Rather than “Gaps in the annotation” I would rather refer to “additional gene models lacking evidence at the transcriptome level”

“was evaluated against known [metazoan|plant|fungal] proteins.” Is unclear

“Low-quality transcript models, such as those showing evidence of fragmented ORFs, were removed. In cases where RNA-seq data were fragmented or absent, homology data were prioritised, favouring longer transcripts with strong intron support from short-read data.” I am not sure about what the authors mean by this. First, the definition of “low quality transcript models” is rather vague. Second, I don’t see how “strong intron support” could be gained from short read data if RNA-seq data was absent, since these two facts are incompatible: the lack of RNA-seq evidence determines the lack of RNA-seq reads that could support exon-exon junctions. This needs to be explained much better than this.

I guess that reads k-mers completeness was estimated using the Illumina library only. This should be explicitly stated.

It is not immediately clear whether and where gene annotations were made publicly available (they should be).

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Invertebrate genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 24 June 2026

https://doi.org/10.21956/openreseurope.25061.r75561

© 2026 Huang X. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Xiao-Chen Huang
Nanchang University, Nanchang, China

This Data Note reports the first chromosome-level reference genome for *Phorcus turbinatus*, a biomarker species for water quality along the Mediterranean coast.

The study employed a combined ONT long-read and Illumina short-read sequencing strategy and achieved chromosome-level scaffolding via whole-genome alignment (WGA) to the sister species *Phorcus lineatus* in the absence of Hi-C data. The final assembly (0.95 Gb) demonstrates high continuity and completeness (BUSCO 99.6%). Data have been deposited in INSDC/ENA/Ensembl with thorough documentation, and the methodology follows the standardized ERGA-BGE pipeline. This resource has practical value for environmental monitoring and evolutionary genomics. However, several critical issues require clarification.

1, The estimated genome size based on ancestral taxa is 1.91 Gb, yet the final assembly spans only ~0.951 Gb — almost exactly half. This should be stated explicitly in the abstract and main text. Otherwise, readers may suspect substantial missing sequence.

2. No Hi-C Data. Chromosome assignment relied entirely on WGA to *P. lineatus.* Figure 2 demonstrates several inversions and translocations between the two species. The super-scaffolding of chromosome 17 is supported by terminal telomeric repeats, but structural validation for the remaining chromosomes remains indirect. The authors should discuss these limitations more thoroughly.

3. Five specimens were collected, but it is not explicitly stated which specimen was used for genome assembly, DNA extraction, and RNA extraction. Three BioSample accessions are listed (SAMEA114492921, SAMEA114492923, SAMEA114492925); their correspondence to the five specimens and respective purposes should be clarified.

4. The annotation section mentions "publicly available short-read RNA-seq data," whereas the Methods section describes RNA extraction and library preparation as part of this study. Please clarify whether the RNA-seq data were newly generated here and whether publicly available external datasets were also used. Additionally, transcriptomic data were derived from a single tissue type ("body with a shell"), which may result in incomplete or fragmented annotation of tissue-specific genes.

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?

Partly

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics and Evolution of Freshwater Bivalves

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 22 June 2026

https://doi.org/10.21956/openreseurope.25061.r74860

© 2026 Gomes-dos-Santos A. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

André Gomes-dos-Santos
University of Porto, Porto, Portugal

I thank *Open Research Europe* for the opportunity to review this manuscript.

In the manuscript entitled *"ERGA-BGE reference genome of Phorcus turbinatus - a water quality biomarker and species distributed along the Mediterranean coast"*, Ortiz et al. present the first reference genome assembly for the coastal Mediterranean gastropod species *Phorcus turbinatus*. The genome is sequenced and assembled to the contig level using state-of-the-art sequencing technologies, then scaffolded to the chromosome level using the genome of a closely related species as a reference. The overall quality of the assembly is tested using standard approaches and attests to the good quality of the assembly.

Overall, the manuscript is well written, and the results appear robust. I have only a concern regarding the vague description of the methods used, which may hamper reproducibility. In the text, the authors refer to the software used, but do not specify the parameters, and the overall workflow is not too clear. I would advise the authors to either provide more detail or, alternatively, a pipeline/workflow of the genome assembly and annotation strategy.

Minor suggestions:

The instruction is missing references. Please clarify this.

When first mentioning the name of the species, please provide the name of the author of first described it, i.e, *Phorcus turbinatus* (Born, 1778).

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Partly

Are sufficient details of methods and materials provided to allow replication by others?
No

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Evolutionary Biology and Genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.